ПСМО ФКН ВШЭ, 3 курс, 1 модуль

Задание 3. Проверка гипотез. Регрессия. Байесовские методы.

Прикладная статистика в машинном обучении, осень 2018

Время выдачи задания: 2 декабря (воскресенье).

Срок сдачи: **16 декабря (воскресенье), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x/PYTHON 3.x.

Правила сдачи

Инструкция по отправке:

1. Решения задач следует присылать единым файлом формата `pdf`, набранным в LaTeX, либо в составе ipython-тетрадки в форматах `ipynb` и `html` (присылайте оба формата, т.к. AnyTask из-за высокой загрузки иногда не рендерит тетрадки в формате `ipynb` – а если мы не увидим ваши задачи, мы их не проверим). Отправляйте практические задачи в виде отдельных файлов (ipython-тетрадок или исходных файлов с кодом на языке `python`).

Оценивание и штрафы:

1. Максимально допустимая оценка за работу над основными задачами – 10 баллов.

2. Бонусные баллы (см. конец домашнего задания) и влияют на освобождение от задач на экзамене.

3. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.

4. Задание выполняется каждым студентом индивидуально и независимо от других студентов. «Похожие» решения считаются плагиатом и все студенты (в том числе те, у кого списали) не могут получить за него больше 0 баллов, причем обнуляются и бонусные баллы. Если вы нашли решение какого-то из заданий (или его часть) в открытом источнике, необходимо указать ссылку на этот источник в отдельном блоке в конце вашей работы (скорее всего вы будете не единственным, кто это нашел, поэтому чтобы исключить подозрение в плагиате, необходима ссылка на источник).

Основные задачи (10 баллов)

1. (2 балла) Пусть $X_1, \dots, X_n \sim \text{Poisson}\,(\lambda)$.
   (a) (1 балл) Пусть $\lambda_0 > 0$. Построить критерий Вальда размера α для различения гипотез $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$.
   (b) (1 балл) Пусть $\lambda_0 = 1$, $n = 20$ и $\alpha = 0.05$. Сгенерировать $X_1, \dots, X_n \sim \text{Poisson}\,(\lambda_0)$ и применить критерий Вальда. Повторить эксперимент много раз и подсчитать долю от общего числа случаев, когда гипотеза H_0 была отклонена. Насколько получившаяся доля ошибок I рода оказалась близкой к 0.05?

2. (1 балл) Пусть $X_1, \dots, X_n \sim N\left(\mu, \sigma^2\right)$. Построить тест на основе отношения правдоподобий для для различения гипотез $H_0 : \sigma = \sigma_0$ vs. $H_1 : \sigma \neq \sigma_0$. Сравнить полученный тест с тестом Вальда для различения этих гипотез.

3. (4 балла) Скачать данные о пробеге автомобиля (в милях) на единицу расхода горючего по ссылке https://github.com/artonson/hse-stat-course-2018/blob/master/homework-handouts/passengercarmile txt
   (a) (1 балл) Подогнать простую линейную регрессию к данным, чтобы предсказать значение переменной MPG (miles per gallon) от значений переменной HP (horsepower). Проанализовать полученные результаты, снабдив их графиком, на котором изображена выборка и оцененная регрессионная зависимость
   (b) (1 балл) Повторить эксперимент из предыдущего пункта, но при этом использовать $\log(\text{MPG})$ в качестве отклика регрессии. Сравнить качество подгонки полученной зависимости с качеством подгонки зависимости из предыдущего пункта (по сумме квадратов остатков подгонки исходных значений MPG).

(c) (1 балл) Подогнать к данным множественную линейную регрессию, чтобы предсказать значение переменной MPG от всех остальных переменных. Проанализовать полученные результаты.

(d) (1 балл) Использовать статистику Mallow C_p (см. слайды 33–49 лекции «Регрессия») для того, чтобы выбрать наилучшее подмножество регрессоров. Использовать и прямой и обратный варианты пошагового выбора. Проанализовать полученные результаты.

4. (1 балл) Допустим, что в регрессионной модели $y = \sum_{j=1}^{k} \beta_j x_j + \varepsilon$ шум $\varepsilon \sim N(0, \sigma^2)$ и дисперсия σ^2 – известна. Показать, что модель с наибольшим значением AIC является моделью с наименьшим значением статистики Mallow C_p.

5. (2 балла) Пусть получена выборка $\{x_n\}_{n=1}^N$ из смеси трёх дискретных распределений:

$$p(x_n) = \sum_{k=0}^{2} w_k p_k(x_n),\ \ w \in \Delta^3$$

$$\begin{array}{rccc} & 0 & 1 & 2 \\ p_0: & 0 & \alpha & 1-\alpha \\ p_1: & \beta & 0 & 1-\beta \\ p_2: & \frac{1}{2} & 0 & \frac{1}{2} \end{array}$$

При этом в выборке нулей 10, единиц 1, а двоек 9. Пусть инициализация на все параметры равномерная. Выпишите две итерации EM алгоритма.

Бонусные задачи (4 балла)

1. (2 балла) Пусть $X_1, \dots, X_n$ – i.i.d. наблюдения. Рассмотрим две модели – M_0 и M_1.

$$M_0 : X_1, \dots, X_n \sim N(0, 1),$$

$$M_1 : X_1, \dots, X_n \sim N(\theta, 1), \, \theta \in \mathbb{R}.$$

По сути, критерии типа AIC (см. слайды 33–49 лекции «Регрессия») позволяют рассмотреть проблему выбора между двумя гипотезами $H_0 : \theta = 0$ и $H_1 : \theta \neq 0$ с точки зрения выбора наилучшей модели. Пусть $l_n(\theta)$ – логарифм функции правдоподобия. Значение AIC для модели M_0 составляет $AIC_0 = l_n(0)$, а значение AIC для модели M_1 составляет $AIC_1 = l_n\left(\hat{\theta}\right) - 1$. Допустим, что выбирается модель с наибольшим значением AIC. Пусть J_n обозначает номер выбранной модели

$$J_n = \begin{cases} 0, & \text{если } AIC_0 > AIC_1; \\ 1, & \text{если } AIC_1 > AIC_0. \end{cases}$$

(a) Допустим, что модель M_0 – верная. Необходимо найти

$$\lim_{n\to\infty} \mathrm{P}\left(J_n = 0\right).$$

Также найдите $\lim_{n\to\infty} \mathrm{P}\left(J_n = 0\right)$ при $\theta \neq 0$.

(b) Пусть $\phi_\theta(x)$ обозначает плотность нормального распределения, среднее значение которого равно θ, а дисперсия равна 1. Определим

$$\hat{f}_n(x) = \begin{cases} \phi_0(x), & \text{если } J_n = 0; \\ \phi_{\hat{\theta}}(x), & \text{если } J_n = 1. \end{cases}$$

Если $\theta = 0$, то показать, что $D\left(\phi_0, \hat{f}_n\right) \to 0$ по вероятности при $n \to \infty$, где

$$D\left(f, g\right) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

является расстоянием Кульбака. Показать также, что $D\left(\phi_\theta, \hat{f}_n\right) \to 0$ по вероятности при $n \to \infty$, если $\theta \neq 0$. Таким образом, AIC состоятельно "оценивает" настоящую плотность распределения несмотря на то, что $\lim_{n\to\infty} \mathrm{P}\left(J_n = 0\right) \neq 1$ при $\theta = 0$.

2. (2 балла) Рассмотрим задачу регрессии с помощью гауссовских процессов с ядром:

$$K(x_i, x_j) = \beta^{-1}[i = j] + x_i^T x_j$$

Покажите, что решение такой задачи совпадает с решением задачи гребневой регрессии. *Подсказка:*

$$(X^T X + \lambda I)^{-1} X^T = X^T (X X^T + \lambda)^{-1}$$